This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery

Patrick Ebel, *Graduate Student Member, IEEE*, Andrea Meraner, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— The majority of optical observations acquired via spaceborne Earth imagery are affected by clouds. While there is numerous prior work on reconstructing cloud-covered information, previous studies are, oftentimes, confined to narrowly defined regions of interest, raising the question of whether an approach can generalize to a diverse set of observations acquired at variable cloud coverage or in different regions and seasons. We target the challenge of generalization by curating a large novel data set for training new cloud removal approaches and evaluate two recently proposed performance metrics of image quality and diversity. Our data set is the first publically available to contain a global sample of coregistered radar and optical observations, cloudy and cloud-free. Based on the observation that cloud coverage varies widely between clear skies and absolute coverage, we propose a novel model that can deal with either extreme and evaluate its performance on our proposed data set. Finally, we demonstrate the superiority of training models on real over synthetic data, underlining the need for a carefully curated data set of real observations. To facilitate future research, our data set is made available online.**

Manuscript received July 9, 2020; revised August 31, 2020; accepted September 14, 2020. This work was supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project "AI4Sentinels–Deep Learning for the Enrichment of Sentinel Satellite Imagery" under Grant FKZ50EE1910. The work of Xiao Xiang Zhu was supported by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme (Acronym: *So2Sat*) under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research," and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future Ai Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond." *(Corresponding author: Xiao Xiang Zhu.)*

Patrick Ebel is with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Andrea Meraner was with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany. He is now with the EUMETSAT European Organisation for the Exploitation of Meteorological Satellites, 64295 Darmstadt, Germany (e-mail: andrea.meraner@eumetsat.int).

Michael Schmitt was with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany. He is now with the Department of Geoinformatics, Munich University of Applied Sciences, 80335 Munich, Germany (e-mail: michael.schmitt@hm.edu).

Xiao Xiang Zhu is with Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.3024744

***Index Terms*— Cloud removal, data fusion, deep learning, generative adversarial network (GAN), optical imagery, synthetic aperture radar (SAR)-optical.**

I. Introduction

ON AVERAGE about 55% of the Earth's land surface is covered by clouds [1], impacting the aim of missions, such as Copernicus, to reliably provide noise-free observations at a high frequency, a prerequisite for applications relying on temporally seamless monitoring of our environment, such as change detection or monitoring [2]–[5]. The need for cloud-free Earth observations, hence, gave rise to a rapidly growing number of cloud removal methods [6]–[12]. While the aforementioned contributions share the common aim of dehazing and declouding optical imagery, the majority of methods are evaluated on narrowly defined and geospatially distinct regions of interest (ROIs). Not only is this specificity posing challenges for a conclusive comparison of methodology but also, furthermore, may cloud-removal performance on a particular ROI poorly indicate performances on other parts of the globe or at different seasons. Moreover, it would be desirable for a cloud removal method to be equally applicable to all regions on Earth, at any season. This generalizability would allow for large-scale Earth observation without the need for costly redesigning or retraining for each individual scene that a cloud removal method is meant to be applied to.

This concern is sustained by previous analysis demonstrating that landcover statistics differ across continents [13] and cloud-coverage is highly variable depending on meteorological seasonality [1]. A major reason for these issues, which is still remaining open nowadays, is the current lack of available large-scale data sets for both training and testing of modern cloud removal approaches. In this work, we address this issue by curating and releasing a novel large-scale data set for cloud removal containing over 100 000 samples from over 100 ROIs distributed over all continents and meteorological seasons of the globe. Especially, we address the challenge of cloud removal in observations from Copernicus mission's Sentinel-2 (S2) satellites. While optical imagery is affected by bad weather conditions and lack of daylight, sensors based on synthetic aperture radar (SAR) as mounted on Sentinel-1 (S1) satellites are not [14] and, thus, provide a valuable source of complementary information. Recent advances in cloud removal combine multimodal data with deep neural networks recovering the affected areas [6], [7], [12], [15].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

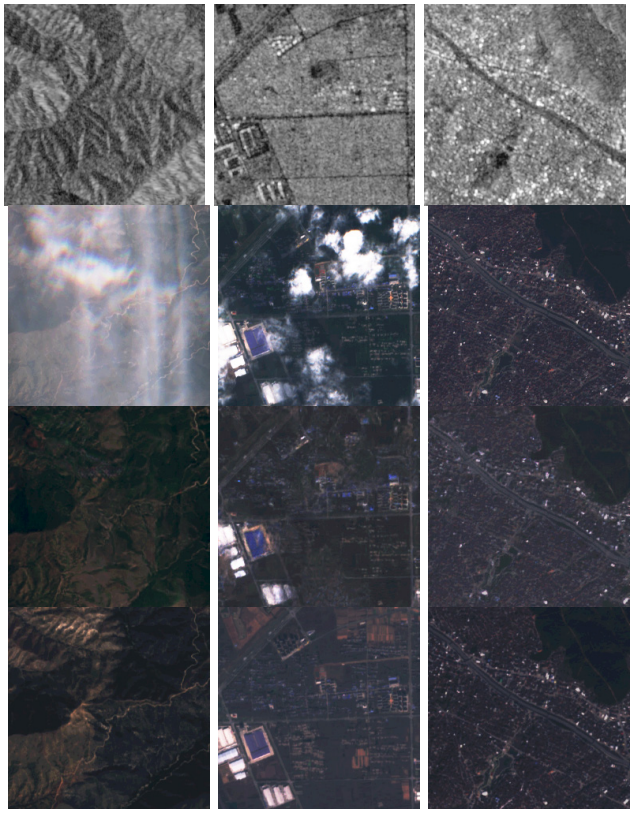


Fig. 1. Exemplary raw data and declouded images. Rows: S1 data (in grayscale), S2 data (in RGB), predicted $\hat{S}2$ data, and cloud-free (target) S2 data. Columns: three different samples. The outcomes show that our model learns to preserve optical data of cloudless areas while replacing cloudy regions by the translation from the SAR domain.

However, many networks are trained on synthetic data or on real data while making strong assumptions on the type and amount of cloud coverage. Moreover, the majority of methods do not explicitly model the amount of cloud coverage and treat each pixel similarly, thereby making unneeded changes to cloud-free areas.

In this work, we address the problem of cloud removal in optical data by means of SAR-optical data fusion, as illustrated in Fig. 1. To redeem the current lack of sufficiently sized and heterogeneous Earth observation data for cloud removal, we release a novel large-scale global data set of coregistered optical cloudy, cloud-free, and SAR observations to train and test the declouding methods. Our data set consists of over 100 000 samples, allowing the training of large models for cloud removal and capturing a diverse range of observations from all continents and meteorological seasons. In addition, we propose a novel generative architecture that reaches competitive performance, as evidenced by two very recently proposed metrics of generated image goodness and diversity. Finally, we show that synthetic data utilized in previous studies are a poor substitute for real cloud coverage data, underpinning the needs for the novel data set proposed in our work.

A. Related Work

The first deep neural architecture to reconstruct cloud-covered images combined near-infrared and red-green-blue (RGB) bandwidth optical imagery by means of a conditional generative adversarial network (GAN) [6], motivated by infrared bandwidth being to a lesser extent impacted by cloud coverage. Subsequent studies replaced the infrared input with SAR observations [7], [15] due to SAR microwaves not being affected by clouds at all [14]. While the early works of [6] and [7] provide a proof-of-concept solely on synthetic data of simulated Perlin noise [16], the networks of [8] and [15] were first to demonstrate performances on real-world data, though focusing primarily on the removal of filmy clouds. Comparable to these studies, we investigate the benefits of SAR-optical data fusion for cloud removal. Unlike the prior work, we address declouding on a carefully curated data set of real imagery sampled over all continents and meteorological seasons, relying neither on synthetic data nor making any strong assumptions about the type and percentage of cloud coverage. Building on the previous studies, the models of [8] and [17] replace the conditional GAN by a cycle-consistent architecture [18], relaxing the preceding models' requirements for pixelwise corresponding training data pairs. While [8] relies solely on cloudy optical input data at inference time, only SAR observations are utilized in [17]. Similar to these two networks, the model that we propose uses a cycle-consistent GAN architecture. We combine cloudy optical with SAR observations and extend on the previous models by incorporating a focus on local reconstruction of cloud-covered areas. This is in line with very recent work [12], [19] that proposed an auxiliary loss term to encourage the model reconstructing information of cloud-covered areas in particular. The network of [12] is noteworthy for two reasons: first, for departing from the previous generative architectures by using a residual network (ResNet) [20] trained supervisedly on a globally sampled data set of paired data; second, for adding a term to the local reconstruction loss that explicitly penalizes the model for modifying off-cloud pixels. Comparable to [12], our network explicitly models cloud coverage and minimizes changes to cloud-free areas. Unlike the model of [12], our architecture follows that of cycle-consistent GAN and has the advantage of not requiring pixelwise correspondences between cloudy and noncloudy optical training data, thereby also allowing for training or fine-tuning on data where such a requirement may not be met. Complementary to the SAR-optical data fusion approach to cloud removal, recent contributions proposed integrating information of repeated observations over time [10], [11]. The work indicates promising results but trades temporal resolution for obtaining a single cloud-free observation, whereas our approach predicts one cloud-free output per cloudy input image and, thus, allows for sequence-to-sequence translation. Moreover, current multitemporal approaches make strong assumptions about the maximum permissible amount of cloud-coverage affecting individual images in the input time series, which is required to be no more than 25% or 50% of cloud coverage for the method of [10] and 10%–30% in the work of [11]. Our curated data sets evidence that such strict requirements on the percentage of cloudiness may, oftentimes, not be met in practice. Consequently, our model makes no assumptions on the maximum amount of tolerable cloud coverage per observation and can gracefully deal with

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

samples ranging from cloud-free to widely obscured skies due to minimizing changes to cloud-free pixels and using SAR observations unaffected by clouds.

II. Methods

We propose a novel model to recover cloud-occluded information in optical imagery. Our network explicitly processes a continuous-valued mask of cloud coverage computed on the fly, as described in Section II-A, to preserve cloud-free pixels while making data-driven adjustments to cloudy areas. The continues-valued assignment of each pixel in the processed cloud mask can be interpreted as the likelihood of the pixel being cloud-covered according to the cloud detector algorithm of [21]. Our model explicitly processing cloud coverage information is in contrast to previous generative architectures that are agnostic to cloud-coverage [6], [8] and networks that only utilize binary cloud mask information [12] as opposed to more fine-grained continuous-valued masks proposed in this work. A cycle-consistent generative architecture detailed in Section II-B allows for training without the need for coregistered cloudy and noncloudy observations of strict pixelwise one-to-one correspondences compared with earlier approaches that required strict pixelwise alignments [7], [15]. We adapt the architecture to integrate SAR with optical observations and propose a new auxiliary cloud map regression loss that enforces sparse reconstructions to minimize modification on cloud-free areas, as described in Section II-C.

A. Cloud Detection and Mask Computation

To evaluate the cloud coverage statistics of our collected data set and model cloud coverage explicitly while reconstructing cloud-covered information, we compute cloud probability masks m. The masks m are computed online for each cloudy optical image and contain continuous pixel values within [0, 1], indicating, for a given pixel, its probability of being cloud-covered. We compute m via the classifier s2cloudless of [21], which demonstrated cloud detection accuracies on par with the multitemporal classifier MAJA [22], running on single-shot observations. While s2cloudless originally applies classification to compute a sparsified binary cloud mask, we wish to obtain a continuous-valued cloud map. We, therefore, take the intermediate continuous-valued representation of the pipeline of [21], then apply a high-pass filter to only keep values above 0.5 intensity, and, finally, convolve with a Gaussian kernel of width $\sigma = 2$ to get a smoothed cloud map with pixel values in [0, 1]. We note that m may alternatively be computed by a dedicated deep neural network [23], but our solution is lightweight and, thus, perfect to support methods running on very large data sets, at almost no additional computational cost in either memory or run time. Exemplary samples of cloud probability masks are presented in Appendix A.

B. Architecture

The model proposed in this work follows the architecture of cycle-consistent GAN [18], i.e., we use two generative networks $G_{S1\to S2}$ and $G_{S2\to S1}$ that translate images from the source domain of $S1$ to the target domain of $S2$, and vice versa. Distribution $\dot{S1}$ (or $\dot{S2}$) denotes the target when the generator performs a within-domain identity mapping, preserving the input image's sensor characteristics. For each domain, there exists an associated discriminator network, denoted as D_{S1} and D_{S2}, respectively, classifying whether a given image is a sample from the domain's true distribution S1 (or S2) or from the synthesized distribution $\hat{S1}$ (or $\hat{S2}$). An overview of our model ensemble is given in Fig. 2. While we keep the network $G_{S2\to S1}$ as in the original work, we apply spectral normalization [24] to both discriminators and make adjustments as follows: $G_{S1\to S2}$ receives an image from domain $S1$ as input and is additionally conditioned on the corresponding cloudy image from $S2$, as well as the cloud probability mask m. For our cloud-removal network, we keep the encoder–decoder architecture of the generator but add a long-skip connection [20] such that the output is given by

$$\hat{S2} = G_{S1\to S2}(\cdot) = \tanh(S2 + S2_{\text{res}})$$

where $S2_{\text{res}}$ denotes the residual mapping learned by the generator. To demodulate the effects of the output nonlinearity on the long-skipped pixels, the inverse hyperbolic tangent is applied to the cloudy input image from $S2$ before the residual mapping. Furthermore, we insert a regression layer taking the residual maps $S2_{\text{res}}$ as input and returning a prediction $\hat{m}$ of the cloud map m. The purpose of the regressor is to enforce a meaningful relation between the learned $S2_{\text{res}}$ and the conditioning m, making the residual maps sparse. Here, sparseness refers to the residual maps being (close to) zero over noncloudy areas, as opposed to having widespread small values, which would indicate many unneeded changes made to cloud-free pixels. We enforce sparseness of the residual maps by formulating an L1 loss on the cloud mask regression, as defined in Section II-A. The loss term effectively acts as a regularizer on changes made to noncloudy areas, penalizing unnecessary adjustments. The regression layer consists of a [3 × 3] convolutional kernel mapping the generated 3-D image to a single-channel map and, thus, adds little to the overall number of learnable parameters. The architecture of generator $G_{S1\to S2}$ is depicted in Fig. 3, and the details on its parameterization are provided in Table I. Discriminator D_{S2} is well-conditioned on the cloud probability maps m. Importantly, we forward the (unpaired) noncloudy optical images to the discriminator D_{S2}, which learns the noncloudy patchwise statistics and, thus, implicitly forces $G_{S1\to S2}$ to synthesize cloud-free images. In sum, our main contribution with respect to architectural changes is twofold. First, we adjusted the generator predicting cloud-free optical images to learn a residual mapping by introducing a long-skip connection forwarding optical information, removing the previous need to reconstruct (even cloud-free) pixels from scratch. Second, our generator learns to constrain modifications to cloud-covered pixels while keeping clear areas unchanged, which is encouraged by introducing a novel layer regressing the cloud coverage map by the learned residual map.

C. Losses

We adjust the losses such that regions regressed as cloud-free in map m remain untouched, while cloudy areas are

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

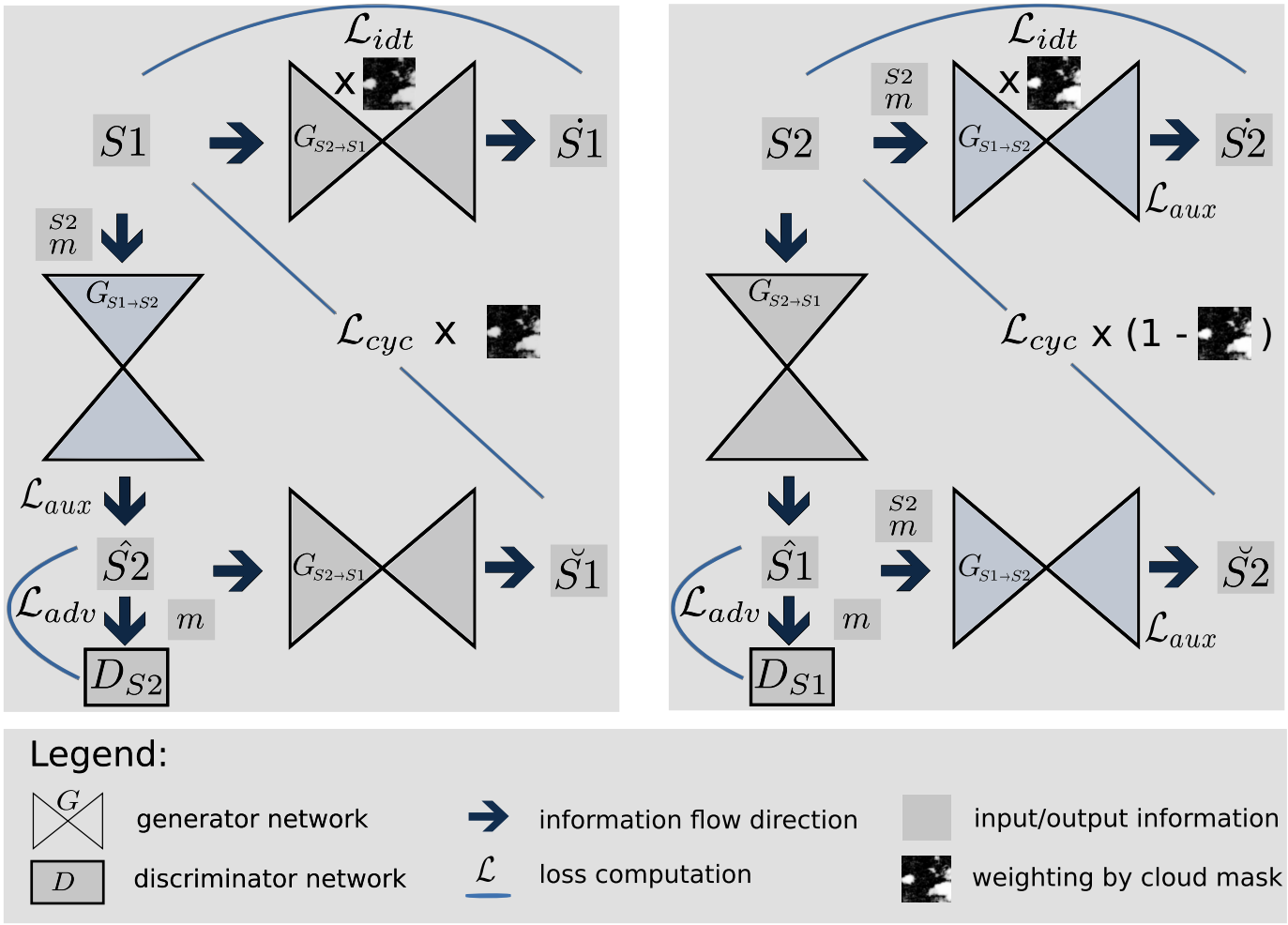


Fig. 2. Overview of our model ensemble based on cycle-consistent GANs [18]. The model consists of two generative networks $G_{S1\to S2}$ and $G_{S2\to S1}$ that translate images from the source domain of $S1$ to the target domain of $S2$, and vice versa. Distribution $\dot{S}1$ (or $\dot{S}2$) denotes the target when the generator performs a within-domain identity mapping, preserving the input image's sensor characteristics. For each domain, there exists an associated discriminator network, denoted as D_{S1} and D_{S2}, respectively, classifying whether a given image is a sample from the domain's true distribution S1 (or S2) or from the synthesized distribution $\hat{S}1$ (or $\hat{S}2$). The network architectures are as in [18]—except for the generator $G_{S1\to S2}$, which is modified as detailed in the main text and in Fig. 3. The losses $\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{cyc}}$, $\mathcal{L}_{\text{idt}}$, and $\mathcal{L}_{\text{aux}}$ are defined in Section II-C.

recovered given the information from domain $S1$. The losses minimized by the generators are

$$\begin{aligned}
\mathcal{L}_{\text{adv}} &= (D_{S1}(\hat{S1}) - 1)^2 + (D_{S2}(\hat{S2}) - 1)^2 \\
\mathcal{L}_{\text{cyc}} &= \|m \cdot (S1 - \breve{S1})\|_1 + \|(1-m) \cdot (S2 - \breve{S2})\|_1 \\
\mathcal{L}_{\text{idt}} &= \|m \cdot (S1 - \dot{S1})\|_1 + \|m \cdot (S2 - \dot{S2})\|_1 \\
\mathcal{L}_{\text{aux}} &= \|(1-m) \cdot (m - \hat{m})\|_1 \\
\mathcal{L}_{\text{all}} &= \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{idt}}\mathcal{L}_{\text{idt}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}
\end{aligned}$$

where $\lambda_{\text{adv}} = 5.0$, $\lambda_{\text{cyc}} = 10.0$, $\lambda_{\text{idt}} = 1.0$, and $\lambda_{\text{aux}} = 10.0$ are the hyperparameters to linearly combine the individual losses within $\mathcal{L}_{\text{all}}$. The loss weightings are set similar to those in [18], with minor adjustments made manually. $\mathcal{L}_{\text{adv}}$ is the adversarial loss originally proposed in LSGAN [25], implementing a least-squares error function on the classifications of the discriminators D_{S1} and D_{S2}. $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{idt}}$ are introduced in [18] but weighted pixelwise with the cloud map m. The purpose of the cycle-consistent loss $\mathcal{L}_{\text{cyc}}$ is to regularizing the mapping $S1 \to S2$ by requiring $S2 \to S1$ being able to reconstruct the original input again (likewise for the direction $S2 \to S1 \to S2$), constraining the potential mappings between both domains. The idea behind $\mathcal{L}_{\text{idt}}$ is to motivate generators to perform an identity mapping and limit unneeded changes in case the provided input is a sample of the target domain. $\mathcal{L}_{\text{aux}}$ is the loss associated with the cloud map regression in $G_{S1\to S2}$, introduced to enforce sparseness of the learned residual feature maps $S2_{\text{res}}$ such that the noncloudy pixels of $S2$ experience little to no adjustments. Our modified generator architecture, the usage of probabilistic cloud maps, and the adjusted losses are showcased in context of a cycle-consistent GAN ensemble, but we remark that they may as well be used within alternative models, such as conditional GAN [26] or ResNet architectures [20].

III. Experiments and Analysis

A. Data

To conduct our experiments, we gather a novel large-scale data set called SEN12MS-CR for cloud removal. For this purpose, we build upon the openly available SEN12MS data set [27] of globally sampled coregistered $S1$ plus cloud-free $S2$ patches and complement the data set with coregistered

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

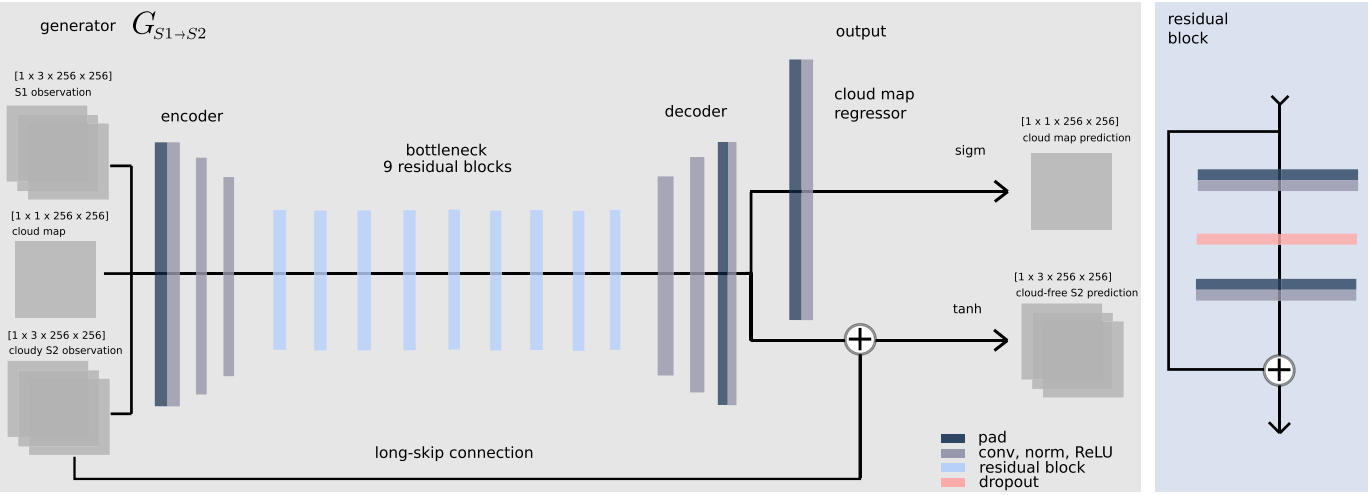


Fig. 3. Detailed architecture of the generator $G_{S1\rightarrow S2}$ of Fig. 2. The generator receives $S1$, m, and $S2$ as input, the latter of which is long-skip forwarded and modified by the learned residual map $S2_{\text{res}}$. The result is passed via a nonlinearity as input to the next network, or treated as output. In parallel, $S2_{\text{res}}$ is regressing m to enforce sparseness of the residual map.

TABLE I

ARCHITECTURE OF OUR GENERATOR $G_{S1\rightarrow S2}$. THE ARCHITECTURE IS DIVIDED INTO FOUR COMPONENTS, AS ILLUSTRATED IN FIG. 3, AND INFORMATION FLOW IS FROM LEFT TO RIGHT ACROSS COMPONENTS AND TOP TO BOTTOM WITHIN COMPONENTS. SYMBOLS: R (RELU), N (INSTANCE NORMALIZATION), C (CONVOLUTION), AND T (TRANSPOSED CONVOLUTION). FOR (TRANSPOSED) CONVOLUTION, THE PARAMETERIZATION IS (KERNEL HEIGHT × KERNEL WIDTH, NUMBER OF FILTERS, STRIDE, AND PADDING SIZE). THE ARCHITECTURE OF GENERATOR $G_{S2\rightarrow S1}$ IS SIMILAR TO THE 9-RESNET BLOCK GENERATOR IN [18], AND THE TWO DISCRIMINATORS ARE KEPT AS THE PATCHGAN DISCRIMINATORS IN [18]

encoder	bottleneck	decoder	output
R(N(C(3 × 3, 64, 1, 1)))	R(N(C(3 × 3, 256, 1, 1)))	R(N(T(3 × 3, 256, 2, 1)))	sigmoid(C(3 × 3, 1, 1, 1))
R(N(C(3 × 3, 128, 2, 1)))	9 × dropout(0.5)	R(N(T(3 × 3, 128, 2, 1)))	tanh(C(3 × 3, 3, 1, 1))
R(N(C(3 × 3, 256, 2, 1)))	R(N(C(3 × 3, 256, 1, 1)))		

cloudy images close in time to the original observations. SEN12MS-CR consists of 169 nonoverlapping ROIs evenly distributed over all continents and meteorological seasons. The ROI has an average size of approximately 52 × 40 km^2 ground coverage, corresponding to complete-scene images of about 5200 × 4000 pixels. Each complete-scene image is checked manually to ensure freedom of noise and artifacts. The cloud-free optical images of four exemplary ROI observed in four different meteorological seasons are depicted in Fig. 4 to highlight the heterogeneity of landcover captured by SEN12MS-CR. Each scene in the data set is subsequently translated into Universal Transverse Mercator coordinate system and then partitioned into patches of size 256 × 256 pixels with a spatial overlap of 50% between neighboring patches, yielding an average of over 700 patches per ROI. Each patch consists of a triplet of orthorectified, georeferenced cloudy, and cloud-free 13-band multispectral Sentinel-2 images, as well as the correspondent Sentinel-1 image (see Fig. 1 for the examples of SAR, cloud-free, and cloudy optical patch triplets). Paired images of the three modalities were acquired within the same meteorological season to limit surface changes. The Sentinel-2 data are from the Level-1C top-of-atmosphere reflectance product. Finally, each patch triples is automatically controlled for potential imaging artifacts, and exclusively, artifact-free patches are preserved to constitute the final cleaned-up version of SEN12MS-CR.

Evaluating the cloudiness of each patch with the algorithm of [21], as described in Section II-A, yields a mean cloud coverage of circa 47.93% ± 36.08%, i.e., about half of all the optical images' information is affected by clouds and the amount of coverage varies considerably. This amount of coverage is notably close to the approximately 55% of global cloud fraction over land that has previously been observed empirically [1]. The distribution of cloud coverage is shown in Fig. 5 and is relatively uniform over the entire domain, with slightly more samples showing (almost) no clouds or being entirely cloud-covered. Note that the computed cloud probability masks are not used to filter any observations or actively guide the data set creation in any manner, and they are solely used *post hoc* to quantify the distribution of cloudiness. For the sake of comparability across models in our experiments and for further studies, we define a train split and a split of hold-out data, which is reserved for the purpose of testing. The train split consists of 114 325 patches sampled uniformly across all continents and seasons and is open to be entirely used for training or in parts for training and validating. The test split consists of 7893 geospatially separated images sampled from ten different ROI distributed across all continents and meteorological seasons, capturing a heterogeneous subset of data.

B. Experiments and Results

A total of three experiments are conducted. First, we train our network and extend it by adding supervised losses for the model to benefit of paired noncloudy and cloudy optical observations in our data set at training time. We systematically

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

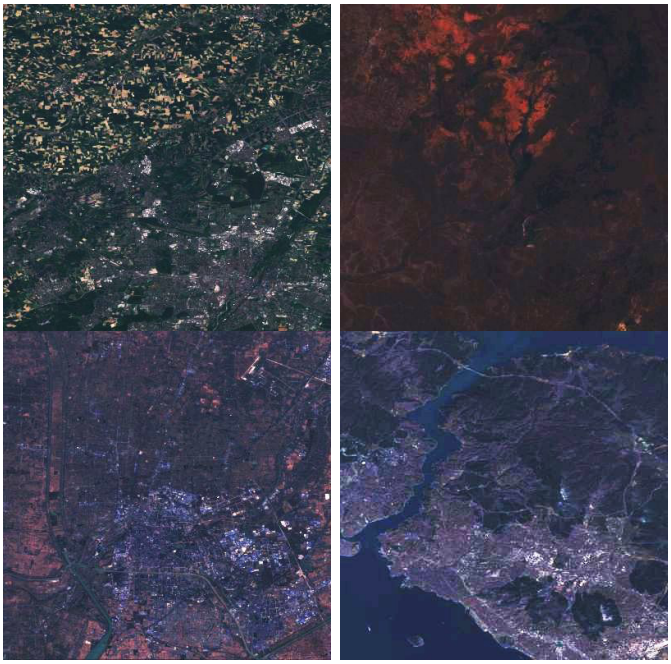


Fig. 4. Cloudless S2 imagery of four exemplary ROI, illustrating the diversity of SEN12MS-CR. The four different scenes are of four different meteorological seasons from the test split of the data set. On average, an ROI is split into over 700 patch samples, each observation of size 256 × 256 pixels.

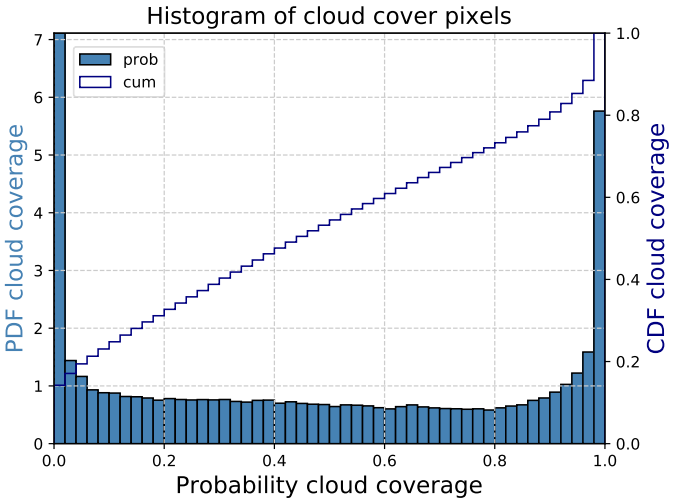


Fig. 5. Statistics of cloud coverage of SEN12MS-CR. On average, approximately 50% of occlusion is observed. The empirical distribution of cloud coverage is relatively uniform and ranges from cloud-free views to total occlusion.

vary the amount of available supervision to investigate its effects on model performance. Second, we evaluate it against a set of baseline models. Third, we retrain the architectures from the previous experiment on synthetic data of generated cloudy observations and evaluate them on real data in order to quantify to which extent models trained on simulated data are capable to generalize to real-world scenarios. To the best of our knowledge, neither of these experiments has previously been conducted in depth. All experiments were conducted on a machine of 8 Intel Core i7-8700 CPU @ 3.20-GHz processors, 16 GB of DIMM DDR4 Synchronous 2667-MHz RAM, and an NVIDIA GeForce RTX 2080, running Ubuntu 18.04. Computation clock time for the training procedure may vary according to the overall task load but is estimated to be about seven days for model ours-0 and about 10–12 days for model ours-100.

1) Metrics to Quantify the Goodness of Cloud Removal: In order to evaluate model performances quantitatively, we utilize the recently developed metrics of improved precision and recall [28], as proposed in the context of generative modeling and improving on previous metrics, such as Inception score or Frechét Inception distance [29], [30]. Improved precision and recall are measures of goodness quantifying similarities between two sets of images in a high-dimensional feature embedding space. Precision is a metric of sample quality, assessing the fraction of generated images that are plausible in the context of the target data distribution. In our context, a generated image is plausible if its high-dimensional feature embedding is sufficiently close to the high-dimensional feature embedding of a cloud-free target image. The distance between both embeddings is sufficiently small if there is no fixed number of neighbors closer to the target embedding than the query embedding. For the formalities behind this metric and motivation of the chosen parameterization, please see Appendix B. Recall measures the diversity in generated images and the extent to which the distribution of target data is covered. Analogous to the metric of precision, a target image is recalled if its high-dimensional feature embedding is sufficiently close to the high-dimensional feature embedding of a generated cloud-free image. Note that this allows interpreting recall as a measure of generated image diversity as the metric can score high only if the generated samples are spread out in the feature embedding's space and provide sufficient coverage of the distribution of target images, capturing the heterogeneity of the target images. To summarize, in the context of our data set of Section III-A, precision specifies the closeness of cloud-recovered information to its cloud-free counterpart, whereas recall captures how well the declouded images capture the heterogeneity of the test data (e.g., its diversity in land-cover and seasonality).

While we emphasize the benefit of both measures to disentangle image quality and image heterogeneity, we also define the F1 score as

$$\mathrm{F1}(X, Y) = 2 \cdot \frac{\mathrm{PR}(X, Y) \cdot \mathrm{RC}(X, Y)}{\mathrm{PR}(X, Y) + \mathrm{RC}(X, Y)}$$

where X and Y are sets of images to be compared, and PR and RC denote the functions of precision and recall, respectively. In contrast to the first two experiments, the generation of synthetic data in the third experiment guarantees a one-to-one pixelwise correspondence between cloudy and ground-truth cloud-free images (i.e., perfect coregistration, no atmospheric disturbances other than the simulated noise, control for no landcover, and daylight changes between both observations), ensuring that pixelwise metrics are well-defined. Therefore, complementary to the previous measures of goodness, we additionally assess performances on synthetic data in the third experiment by means of mean absolute error

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

(MAE), root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [31], and spectral angle mapper (SAM) [32], as given by

$$\mathrm{MAE}(x, y) = \frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} |x_{c,h,w} - y_{c,h,w}|$$

$$\mathrm{RMSE}(x, y) = \sqrt{\frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} (x_{c,h,w} - y_{c,h,w})^2}$$

$$\mathrm{PSNR}(x, y) = 20 \cdot \log_{10} \left(\frac{1}{\mathrm{RMSE}(x, y)} \right)$$

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x \mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x + \mu_y + \epsilon_1)(\sigma_x + \sigma_y + \epsilon_2)}$$

$$\mathrm{SAM}(x, y) = \cos^{-1} \times \left(\frac{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w} \cdot y_{c,h,w}}{\sqrt{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w}^2 \cdot \sum_{c=h=w=1}^{C,H,W} y_{c,h,w}^2}} \right)$$

where x and y are images to be compared with pixel-values $x_{c,h,w}, y_{c,h,w} \in [0, 1]$, dimensions $C = 3$, $H = W = 256$, means μ_x, μ_y, standard deviations σ_x, σ_y, covariance σ_{xy}, and small numbers ϵ_1 and ϵ_2 to stabilize the computation. MAE and RMSE both are pixel-level metrics quantifying the mean deviation between target and predicted images in absolute terms and units of the measure of interest, respectively. PSNR is an imagewise metric to measure how good of a reconstruction in terms of signal-to-noise ratio a recovered image is to a clear target image. SSIM is a second imagewise metric, quantifying the structural differences between the target and predicted images. It is designed to capture perceived change in structural information between two given images, as well as differences in luminance and contrast [31]. The SAM metric is an imagewise measure, quantifying the spectral angle between two images, measuring their similarity in terms of rotations in the space of spectral bands [32]. Further technical information with respect to the metrics utilized in our experiments to quantify goodness of predictions is provided in Appendix B.

2) Quantifying the Benefits of Paired Data: First, we train the architecture described in Section II without using any pixelwise correspondences, as in a manner conventional for cycle-consistent GAN. For our generative model, we consider the VV and VH channels of images from the S1 domain and add a third mean (VV and VH) channel to satisfy the dimension-preservation requirement of cycle-consistent architectures. For images from the S2 domain, all multispectral information is used when computing cloud probability maps, while the S1–S2 mapping uses exclusively the three RGB channels. All images are value-clipped and rescaled to contain values within $[-1, 1]$, while the cloud probability map values are within $[0, 1]$. Value-clipping is within ranges $[-25; 0]$ and $[0; 10\,000]$ for S1 and S2, respectively. Notably, before training, we perform an imagewise shuffling of the optical data of paired cloudy and cloud-free observations to remove the pixelwise correspondences satisfied when cloudy and cloud-free patches would be available as sorted tuples. That is, the optical cloudy and noncloudy patches presented at one training step may be no longer strictly aligned or could reflect differences in landcover and atmosphere, reflecting practical challenges commonly encountered when gathering data for remote sensing applications. We train our network on a 10 000 images multiregion subset of the training split introduced in Section III-A. Network weights w are initialized by sampling from a Gaussian distribution $w \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.02)$. The optimizer and the hyperparameters for the optimizer and loss weightings are set as in [18]: We use ADAM with an initial learning rate $\epsilon_{\mathrm{lr}} = 0.0002$, momentum parameters $\beta = (0.5, 0.999)$ for computing sliding averages of the gradients, and their squares and a small constant of 10^{-8} added to the denominator to ensure numerical stability of the optimizer. Instance normalization [33] is applied to the generators as in the original architecture [18], with adjustments detailed in Fig. 3 and Table I. Spectral normalization [24] is applied to the discriminators as in [34] in order to prevent mode collapse during training [35]. The networks are trained for $n_{\mathrm{iter}} = 50$ epochs at the initial learning rate of ϵ_{lr} and then for another $n_{\mathrm{decay}} = 25$ epochs with a multiplicative learning rate decay given by $\mathrm{lr}_{\mathrm{decay}}(n_{\mathrm{current}}) = 1.0 - \max(0, 1 + n_{\mathrm{current}} - n_{\mathrm{iter}})/(n_{\mathrm{decay}} + 1)$, where n_{current} denotes the current epoch number. The gentle learning rate decay over a long period of epochs serves to ensure a well-behaved optimization process during training [18], [35]. All our generator networks are trained on center-cropped 200×200 px^2 patches but tested on full-sized 256×256 pixels patches of the hold-out split, as the generator architecture is fully convolutional. As proposed in [36] and implemented in [18], we maintain two pools of the last 50 generated images to update the discriminators with a random sample from the respective image buffers such that oscillations during training are reduced [18], [35]. Representative qualitative outcomes are depicted in Fig. 1. The results highlight that our model can reconstruct cloud-covered areas while preserving information that is not obscured. A quantitative evaluation of the described model (ours-0) is given in Table II.

TABLE II
EFFECT OF PERCENTAGE OF PAIRED TRAINED DATA ON PERFORMANCE OF CLOUD REMOVAL MODEL. THE MORE THE PAIRED TRAINING DATA IS AVAILABLE, THE BETTER THE RESULTING PERFORMANCES

% paired	precision	recall	F1 score
0 (ours-0)	0.560	0.491	0.523
10	0.559	0.499	0.527
20	0.560	0.506	0.532
50	0.562	0.528	0.544
100 (ours-100)	**0.564**	**0.551**	**0.557**

Second, we retrain the model, as described earlier, but on paired cloudy–cloudless optical observations in order to assess the benefits of paired training data, as provided by our data set. To let the cycle-consistent architecture described in Section II benefit of paired training data, we combine the losses defined in Section II-C with cost functions defined on paired images: first, a pixelwise L1 loss penalizing prediction errors between generated and paired target images as in [37]; second, perceptual losses for features and style [38], as evaluated on the features extracted at ReLU layers 11, 20, and 29 of an auxiliary pretrained VGG16 network [39]. We retrain our

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

network with these losses and systematically vary the percent of paired cloudy and cloud-free optical data available. The paired patches are equally spaced across the training split at the beginning of the training procedure, and patch pairings are fixed across epochs. During training, the presentation of paired and unpaired samples occurs in random order. Table II shows the different models' performances. The base model trained on unpaired data (ours-0) performs worst, while the model fully trained on paired samples (ours-100) achieves the best performances. In general, the more paired samples are available the better the model performs.

3) Model Ablation Experiment: To put the results of the previous experiment into perspective and further evaluate the factors benefiting the robust reconstruction of cloud-covered information, we conduct an ablation study. Especially, we investigate the effectiveness of the novel cloud detection mechanism explained in Section II-A and the local cloud-sensitive loss introduced in Section II-C. For this purpose, we retrain the model ours-0, as described in Section II, but omit the cloud-sensitive terms by fixating the values of all pixels in the cloud probability masks m to 1.0. The effect of this is that the ablated model is no longer encouraged to minimize the changes to areas free of cloud coverage, thus potentially resulting in unneeded changes. As additional baselines, we evaluate the goodness of simply using the S1 observations (VV- or VH-polarized), as well as cloud-covered S2 images as predictions and comparing against their cloud-free counterparts. Table III reports the declouding performance of baseline models and our models (0% and 100% paired data from Table II). Our network of 100% paired data performs best in terms of precision and F1 score. The raw S1 and S2 observations perform relatively poorly, except for the cloudy optical images scoring high on image diversity due to random cloud coverage. While it may be useful to consider the raw data as baselines, it is necessary to keep in mind that modalities, such as SAR, maybe at a disadvantage when directly comparing against the cloud-free optical target images.

4) Assessing the Goodness of Synthetic Data: To compensate for the lack of any large-scale data set for cloud removal, previous works simulated the artificial data [6], [7], [10], [40], [41] of synthetic cloudy optical images. This raises the question of the goodness of the simulated observations, i.e., how good of an approximation such simulations are to any real data. In this experiment, we consider the two architectures ours-0 and ours-100 from Table III and retrain them on synthetic data to subsequently evaluate the retrained models on the real test data and assess if performance generalizes to real-world scenarios. Two approaches to generating synthetic data are evaluated.

1) *Perlin:* We generate cloudy imagery via Perlin noise [16] and alpha-blending as in the preceding studies of [6], [7], and [40]. This approach has the limitation of adding Perlin noise to all of the multispectral bandwidths evenly, due to lack of a better physical model of multispectral cloud noise. Since cloud detectors trained on real observations are expected to fail in such a case, we substitute the cloud map of Section II-A by the synthesized alpha-weighted Perlin noise.

TABLE III
CLOUD-REMOVAL PERFORMANCE OF BASELINE METHODS AND OUR MODELS ON TEST SPLIT OF SEN12MS-CR. ROWS S1 VV AND VH REFER TO THE RAW S1 IMAGE, CHANNELS VV AND VH, RESPECTIVELY, COMPARED WITH THE GRAY-SCALE CLOUD-FREE S2 IMAGE. S2 CLOUDY REFERS TO THE RAW CLOUDY S2 IMAGE COMPARED WITH THE RGB CLOUD-FREE S2 IMAGE. ALL MODELS' METRICS BEAT THE LOWER-BOUND PERFORMANCES ESTABLISHED BY THE RAW DATA, EXCEPT ON THE RECALL METRIC. THE FULL MODELS PERFORM BETTER THAN THE ABLATION MODELS WITHOUT THE CLOUD-SENSITIVE LOSS AND CLOUD PROBABILITY MASKS. MODEL OURS-100 PERFORMS BEST IN TERMS OF PRECISION AND F1 SCORE. NOTE THAT THE RESULTS DEPICT A PRONOUNCED TRADEOFF BETWEEN PRECISION AND RECALL, AS ANALYZED, IN DETAIL, IN [28]

model		precision	recall	F1 score
S1	VV	0.000	0.001	0.001
	VH	0.012	0.017	0.014
S2 cloudy		0.161	**0.705**	0.267
ours-0 (no m)		0.181	0.572	0.279
ours-100 (no m)		0.232	0.535	0.323
ours-0		0.560	0.491	0.523
ours-100		**0.564**	0.551	**0.557**

2) *Copy:* We generate cloudy imagery by taking the ground-truth cloud-free optical observations and combine them via alpha-blending with clouded observations as in the approach of [10]. Different from [10], we benefit from our curated data set and alpha-blend paired cloudy–cloudless observations, whereas the prior study mixed the two unrelated images. Moreover, we alpha-blend weighted by the cloud map of Section II-A, whereas the original study alpha-blended via sampled Perlin-noise. We believe that these modifications better preserve the spectral properties of real observations and keep cloud distribution statistics closer to that of real data, as shown in Figs. 6 and 7.

Furthermore, this allows for synthesizing coverage ranging from semitransparent to fully occluded clouds, which would be less straightforward on unpaired observations. Exemplary observations generated by both simulation approaches and empirical observations are presented in Fig. 6.

The outcomes of this experiment are presented in Table IV. For all data simulation approaches, training a network on generated data and, subsequently, evaluating it on synthetic test data are overestimating the performances on the corresponding real test data. This observation holds for both models evaluated in the experiment. The models display a drop in performance when moving from synthetic to real testing data. The drop being considerably smaller in the case of copy–paste data than for Perlin noise data may be due to the copy-pasted data closer resembling the real data and its underlying sta-

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

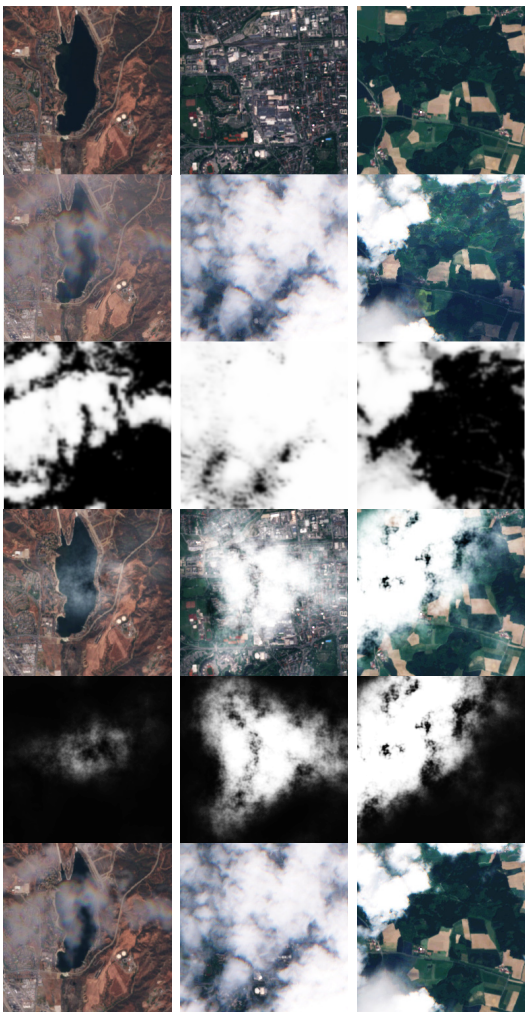


Fig. 6. Exemplary cloud-free, real cloudy, and generated cloudy optical observations. Rows: cloud-free S2 data (plotted in RGB), real cloudy S2 data, real cloud coverage maps (same for copy–paste), Perlin-noise simulated cloudy S2 data, Perlin-noise cloud coverage maps, and copy–paste simulated cloudy S2 data. Columns: three different samples.

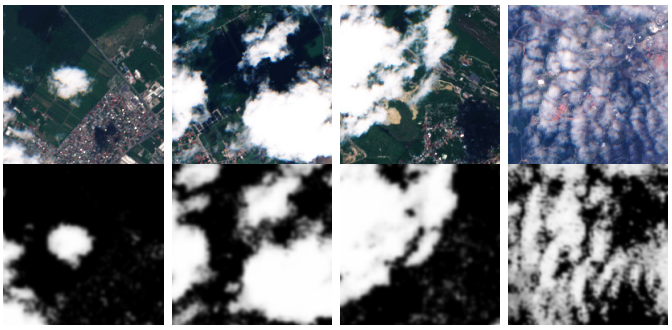


Fig. 7. Exemplary cloudy optical observations and cloud maps. Rows: cloudy S2 data and cloud probability masks. Columns: four different samples.

tistics of cloud coverage and spectral distributions. In this context, it is instructive to investigate spectral distortions by means of SAM, which indicates that models trained and tested on synthetic data are considerably poorer to predict spectral distributions on Perlin-simulated data compared with the copy-pasted observations, which is arguable more alike

TABLE IV
CLOUD-REMOVAL PERFORMANCE OF MODELS OURS-0 AND OURS-100 FROM TABLE III, RETRAINED ON SYNTHETIC CLOUD DATA (EITHER PERLIN-SIMULATED OR COPY-PASTED) AND TESTED ON SYNTHETIC AND REAL DATA. BOTH MODELS, WHEN TRAINED ON SYNTHETIC DATA, PERFORM MUCH BETTER ON SYNTHETIC TEST DATA THAN ON REAL TEST DATA. IMPORTANTLY, THE TEST PERFORMANCE OF MODELS TRAINED ON SYNTHETIC AND TESTED ON REAL DATA IS CONSIDERABLY POORER THAN THAT OF THE SAME ARCHITECTURES TRAINED ON REAL DATA (REPORTED IN TABLE III)

model		ours-0		ours-100	
metric		Perlin	copy	Perlin	copy
MAE		0.045	0.023	0.041	0.017
RMSE		0.067	0.031	0.059	0.023
PSNR		24.75	34.034	25.775	35.802
SSIM		0.803	0.882	0.824	0.904
SAM		27.527	10.626	26.013	9.936
precision	synth	0.155	0.693	0.239	0.692
	real	0.115	0.425	0.168	0.458
recall	synth	0.781	0.851	0.800	0.856
	real	0.624	0.611	0.592	0.586
F1	synth	0.258	0.764	0.368	0.766
	real	0.194	0.501	0.262	0.514

to real data in terms of its spectral properties. The findings in this experiment underline the need for synthetic data to closely capture the properties of real data, yet even when real and synthetic observations may be hardly distinguishable by eye (as the examples shown in Fig. 6), there persist important discrepancies unaccounted for, which hinders the models trained on synthetic sampled to perform equally on real data.

IV. DISCUSSION

The contribution of our work is in providing a large-scale and global data set for cloud removal and developing a new model for recovering cloud-covered information to highlight the data sets benefits. With over 55% of the Earth's land surface covered by clouds [1], the ability to penetrate cloud coverage is of great interest to the remote community in order to obtain continuous and seamless monitoring of our planet. While the focus in this work is on providing the first globally sampled multimodal data set for general-purpose cloud removal, future research should also address the benefits of cloud removal approaches for particular applications common in remote sensing. An example application is in semantic segmentation, which necessitates clear-view observations for accurate land-cover classification. Another, in the context of having consecutive observations over time, would be change or anomaly detection where cloud removal methods may be beneficial particularly for the purpose of early stage detection, which could, otherwise, be delayed in the presence of clouds. A limitation of our proposed cloud removal model is its restriction to work on a subset of the optical observation's spectral bands. While this constraint is required due to the choice of the network architecture as necessitated by our experiments conducted, we are certain that it will be beneficial

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

for future research to consider the full spectral information. To allow for this, our curated global data set is released with all available information for both modalities, including the full spectrum of bands for the optical observations.[1]

V. Conclusion

We demonstrated the declouding of optical imagery by fusing multisensory data, proposed a novel model, and released the, to the best of our knowledge, first global data set combining over a 100 000 paired cloudy, cloud-free, and coregistered SAR sample triplets. Statistical analysis of our data set shows a relatively uniform distribution of cloud coverage, with clear images occurring just as probable as wide and densely occluded ones—indicating the need for flexible cloud removal approaches to potentially handle either case. Our proposed network explicitly models cloud coverage and, thus, learns to retain cloud-free information while as well being able to recover information of areas covered by wide or dense clouds. We evaluated our model on a globally sampled test set and measure the goodness of predictions with recently proposed metrics that capture both prediction quality and coverage of the target distribution. Moreover, we showed that our model benefits from supervised learning on paired training data as provided by our large-scale data set. Finally, we evaluated the goodness of synthetically generated data of cloudy–cloudless image pairs and show that great performance on synthetic data may not necessarily translate to equal performance on real data. Importantly, when testing on real data, the networks trained on real observations consistently outperform models trained on synthetic observations, indicating the existence of properties of the real observations not modeled sufficiently well by the simulated data. This underlines the need for a set of real observations numerous enough to train large models, as provided by the data set released in this work. In further studies, we will address the fusion of multitemporal and multisensory data, combining and comparing across both currently segregated approaches. To support future research and make contributions comparable, we share our global data set of paired cloudy, cloud-free, and coregistered SAR imagery and provide our test data split for benchmarking purposes.

Appendix A
Cloud Detection

We present exemplary cloudy optical observations and cloud maps in Fig. 7. The cloud masks are as predicted by our cloud detection pipeline detailed in Section II-A. The illustrated examples show that our proposed method can reliably detect clouds and provide continuous-valued cloud masks.

Appendix B
Improved Precision and Recall

We provide a definition of improved precision and recall in line with the definitions in [28]. For further details, the interested reader is referred to the original publication.

Definition (Improved Precision and Recall [28]): Let $X_r \sim P_r$ and $X_g \sim P_g$ denote paired samples drawn from the real and generated distributions of cloud-free images, where P_g is the distribution learned by the generator network whose quality is to be assessed. Each sample is mapped via an auxiliary pretrained network M^2 in a high-dimensional feature space to obtain latent representations $\phi_r = M(X_r)$ and $\phi_g = M(X_g)$ such that the two sets of samples are mapped into two feature sets Φ_r and Φ_g. A distribution $P \in \{P_r, P_g\}$ is approximated by computing pairwise distances between feature embeddings of the observed samples $\Phi \in \{\Phi_r, \Phi_g\}$ and, centered at each feature $\phi \in \Phi$, forming a hypersphere with a radius corresponding to the distance to its kth nearest neighbor embedding $N_k(\phi)$. Hence, whether an embedded sample ϕ falls on manifold Φ or not is given via

$$f(\phi, \Phi) = \begin{cases} 1, & \text{if } \exists \phi' \in \Phi : \|\phi - \phi'\| \leq \|\phi' - N_k(\phi')\|_2 \\ 0, & \text{else.} \end{cases}$$

The fraction of samples that fall on the paired distribution's manifold are then defined in [28] as

$$\text{precision}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r)$$

$$\text{recall}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g).$$

We set parameters $|\Phi| = 7893$ corresponding to the size of the test split of SEN12MS-CR and $k = 10$ because every sample has up to 50% overlap with its neighboring samples. This setting removes the paired target itself plus its eight overlapping samples when computing $N_k(\phi)$.

Appendix C
Cloud Coverage Statistics on Test Split

In addition to the cloud coverage statistics on the entire data set, as reported in Section III-A, Fig. 8 provides the empirically observed distribution of cloud coverage on the data sets test split. Even though the histogram of the test split is less smooth than that of the complete data set due to the test split being much smaller, both distributions are considerably alike.

Appendix D
Exemplary Problematic Cases

For the sake of completeness, we discuss cases that we consider challenging for cloud removal approaches, specifically our method, and present exemplary data and predictions of such cases in Fig. 9. We consider the following challenges.

1) Changes in landcover, atmosphere, day time acquisition, or seasonality that may occur between (visible parts of) the cloudy reference image and the cloud-free target optical image. While our data set is curated to minimize such cases by selecting observations that are close in

[1]The SEN12MS-CR data set is shared under the CC-BY 4.0 open access license and available for download provided by the library of the Technical University of Munich (TUM): https://mediatum.ub.tum.de/1554803. This article must be cited when the data set is used for research purposes.

[2]Here, VGG16 [39], with features extracted at the second fully connected layer, as argued for in [42]. Metric evaluation on alternative pretrained networks has shown to provide virtually identical results [28].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

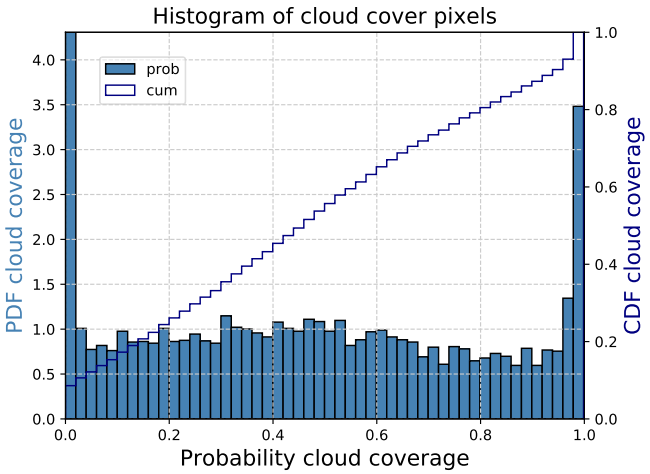


Fig. 8. Statistics of cloud coverage of test split of SEN12MS-CR. As for the statistics on the complete data set, an average of circa 50% of occlusion is observed.

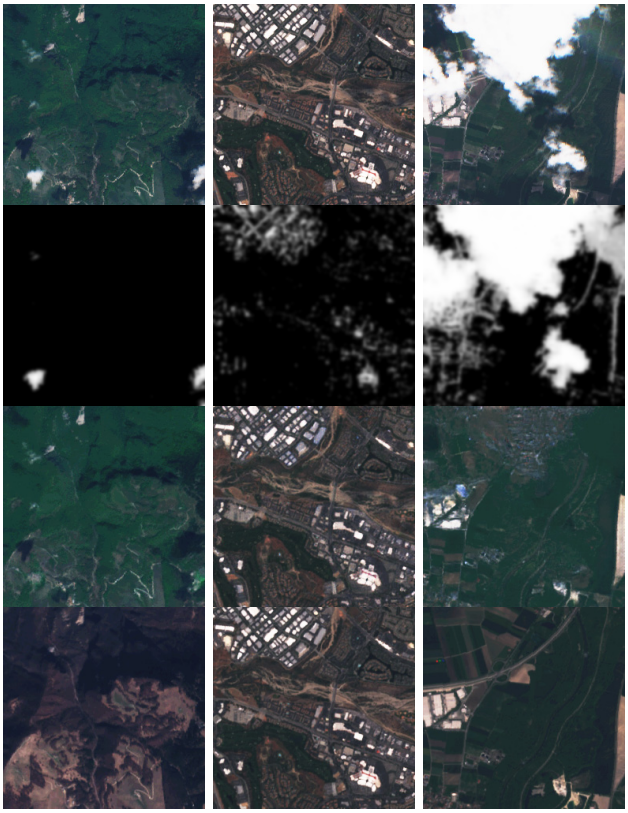


Fig. 9. Exemplary cases posing challenges to our cloud-removal approach. Rows: S2 data (in RGB), predicted cloud map m, predicted $\hat{S2}$ data, and cloud-free (target) S2 data. Columns: three different samples. Reconstructing optical information obscured by clouds is a hard problem. Among the challenges faced by cloud removal approaches may be: 1) overtime changes in landcover, atmosphere, day time acquisition, or seasonality; 2) precise detection of clouds with few misses and false alarms; and 3) correct reconstruction of information fully covered by large and dense clouds.

time, strict ground-truth correspondence is challenging to establish and may only be guaranteed by simulating synthetic data as in experiment III-B4.

2) Precise detection of clouds and accurate cloud masks that minimizes false alarms and misses. With respect to our cloud detection algorithm, there exist cloud masks where, even for completely cloud-free images, pixels are assigned a nonzero (albeit rather low) probability of being cloudy.
3) Correct reconstruction of cloud-covered information. In particular, for the case of complete coverage by large and dense clouds, this is a very challenging problem. We observed the cases where the information reconstructed by our model did not match the target images; for instance, urban-like landcover was predicted in place of agricultural areas.

ACKNOWLEDGMENT

The authors would like to thank ESA and the Copernicus Program for making the Sentinel observations accessed for this submission publicly available. They would also like to thank Lloyd Hughes for having shared his artifact detection preprocessing code with them.

REFERENCES

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[2] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[3] J. R. Jensen *et al.*, *Introductory Digital Image Processing: A Remote Sensing Perspective*, no. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[4] P. Coppin, E. Lambin, I. Jonckheere, and B. Muys, "Digital change detection methods in natural ecosystem monitoring: A review," in *Analysis of Multi-Temporal Remote Sensing Images*. Singapore: World Scientific, 2002, pp. 3–36.

[5] C. E. Woodcock, T. R. Loveland, M. Herold, and M. E. Bauer, "Transitioning from change detection to monitoring with remote sensing: A paradigm shift," *Remote Sens. Environ.*, vol. 238, Mar. 2020, Art. no. 111558.

[6] K. Enomoto *et al.*, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1533–1541. [Online]. Available: https://arxiv.org/abs/1710.04835

[7] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1726–1729. [Online]. Available: https://ieeexplore.ieee.org/document/8519215/

[8] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1772–1775. [Online]. Available: https://ieeexplore.ieee.org/document/8519033/

[9] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1220–1224, Aug. 2019.

[10] M. U. Rafique, H. Blanton, and N. Jacobs, "Weakly supervised fusion of multiple overhead images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1479–1486.

[11] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1796–1805.

[12] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[13] M. A. Friedl *et al.*, "Global land cover mapping from MODIS: Algorithms and early results," *Remote Sens. Environ.*, vol. 83, nos. 1–2, pp. 287–302, Nov. 2002.

[14] R. Bamler, "Principles of synthetic aperture radar," *Surv. Geophys.*, vol. 21, nos. 2–3, pp. 147–157, 2000.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[15] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa, "SAR to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 5–11, Sep. 2018.

[16] K. Perlin, "Improving noise," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 681–682.

[17] M. F. Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial network—Optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251. [Online]. Available: http://ieeexplore.ieee.org/document/8237506/

[19] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," *Remote Sens.*, vol. 12, no. 1, p. 191, Jan. 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] A. Zupanc. (2017). *Improving Cloud Detection With Machine Learning*. Accessed: Oct. 10, 2019. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[22] V. Lonjou *et al.*, "MACCS-ATCOR joint algorithm (MAJA)," *Proc. SPIE*, vol. 10001, Oct. 2016, Art. no. 1000107.

[23] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, Aug. 2019.

[24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[25] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[27] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7. 2019, pp. 153–160, doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

[28] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3929–3938.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] F. A. Kruse *et al.*, "The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data," *AIP Conf.*, vol. 283, no. 1, pp. 192–201, 1993

[33] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," Nov. 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[34] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–26.

[35] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[36] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2107–2116.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, vol. 9906. Cham, Switzerland: Springer, 2016, p. 694–711.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[40] W. Sintarasirikulchai, T. Kasetkasem, T. Isshiki, T. Chanwimaluang, and P. Rakwatin, "A multi-temporal convolutional autoencoder neural network for cloud removal in remote sensing images," in *Proc. 15th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Jul. 2018, pp. 360–363.

[41] D. Tedlek, S. Khoomboon, T. Kasetkasem, T. Chanwimaluang, and I. Kumazawa, "A cloud-contamination removal algorithm by combining image segmentation and level-set-based approaches for remote sensing images," in *Proc. Int. Conf. Embedded Syst. Intell. Technol., Int. Conf. Inf. Commun. Technol. Embedded Syst. (ICESIT-ICICTES)*, May 2018, pp. 1–5.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–35.

Patrick Ebel (Graduate Student Member, IEEE) received the B.Sc. degree in cognitive science from the University of Osnabrück, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is pursuing the Ph.D. degree with the SiPEO Laboratory, Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany.

His research interests include deep learning and its applications in computer vision and to remote sensing data.

Andrea Meraner received the B.Sc. degree in physics and the M.Sc. degree (Hons.) in Earth oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2016 and 2019, respectively.

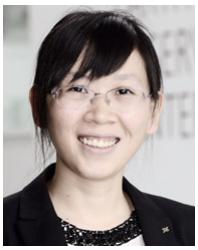
In 2017, he was a Research Assistant with Atmospheric Modeling Group, German Geodetic Research Institute (DGFI-TUM). In 2018, he was with the German Aerospace Center (DLR)–German Remote Sensing Data Center, Weßling, Germany, in the International Ground Segment department. After spending one semester at IIT Mandi, Suran, India, in 2019, he was a Research Assistant with the Signal Processing in Earth Observation (SiPEO) Group, TUM, and the Remote Sensing Technology Institute, DLR, working on deep learning-based cloud removal algorithms for optical satellite imagery. Since October 2019, he has been a Junior Remote Sensing Scientist for optical imagery at EUMETSAT European Organization for the Exploitation of Meteorological Satellites, Darmstadt, Germany, where he is developing algorithms to process and analyze data from current and future geostationary satellite missions.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Michael Schmitt (Senior Member, IEEE) received the Dipl.Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the Habilitation degree in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2020, he has been a Full Professor of applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences, Munich. From 2015 to 2020, he was a Senior Researcher and the Deputy Head at the Professorship for Signal Processing in Earth Observation, TUM. In 2019, he was additionally appointed as an Adjunct Teaching Professor at the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist with the University of Massachusetts at Amherst, Amherst, MA, USA. His research focuses on image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations. In particular, he is interested in remote sensing data fusion with a focus on synthetic aperture radar (SAR) and optical data.

Dr. Schmitt is the Co-Chair of the Working Group "SAR and Microwave Sensing" of the International Society for Photogrammetry and Remote Sensing and the Working Group "Benchmarking" of the IEEE-GRSS Image Analysis and Data Fusion Technical Committee. He frequently serves as a reviewer for a number of renowned international journals and conferences and has received several best reviewer awards. He is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.



Xiao Xiang Zhu (Senior Member, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a co-coordinator of the Munich Data Science Research School. Since 2019, she also heads the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. She is a Professor with the Signal Processing in Earth Observation, TUM, and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.